УДК 32.973.202:07.681

к.т.н. **Бабіч О.М.** (ВІКНУ)
Дух Д.І. (КНУ)
Глухова А.С. (КНУ КіМ)

ОСОБЛИВОСТІ МЕТОДІВ МАШИННОГО НАВЧАННЯ ЩОДО ЇХ ВИКОРИСТАННЯ В ПРОЦЕСІ АНАЛІЗУ АНГЛОМОВНИХ ДЖЕРЕЛ

Досліджено методи машинного навчання для оцінювання емоційної забарвленості текстів англомовних засобів масової інформації. Розглянуто такі методи як наївний метод Байєса, логістична регресія, штучні нейронні мережі. Визначені їх переваги і недоліки щодо використання у вказаній задачі. Так, незважаючи на швидке навчання, Наївний Байєсівський класифікатор, через неможливість знаходження імовірностей значень при класифікації, не є оптимальним алгоритмом для оцінювання емоційної забарвленості англомовних текстів ЗМІ і не може бути достатньо продуктивним. Метод логістичної регресії, у свою чергу, передбачає вивчення квадратичних зв'язків для виконання задачі аналізу емоційної забарвленості англомовних текстів ЗМІ, що збільшить як кількість вхідних параметрів, так і час навчання алгоритму. В ході дослідження надано перевагу методу Штучної нейронної мережі, алгоритми якого дають можливість вивчати найрізноманітніші взаємозв'язки між даними, при цьому не потребуючи значних витрат часу на роботу алгоритму. Аналіз роботи алгоритмів досліджених у статті методів машинного навчання проводився за допомогою методів метрики точності, дерева рішень, AUC-ROC та F-міра.

Ключові слова: інформаційно-аналітична діяльність, засоби масової інформації, машинне навчання, класифікація, функція, алгоритм, нейронна мережа.

Вступ і постановка проблеми. Виклики сучасного інформаційного середовища сприяють розвитку інформаційно-аналітичної діяльності (ІАД), зумовлюючи потребує в удосконаленні обробки даних та в об'єктивізації висновків. Важливим для проведення ІАД є оперативний і поглиблений аналіз іншомовних інформаційних масивів, зокрема текстових повідомлень англомовних засобів масової інформації (ЗМІ).

Один зі шляхів досягнення цього – автоматизація процесу багатовимірної аналітичної обробки повідомлень ЗМІ, що включає інтеграцію процесу обробки даних і динамічної актуалізації вихідних умов у складі системи обробки інформаційного потоку. Для аналізу інформаційного контенту і прогнозування його розвитку в Інтернет-просторі у [1,2] запропоновано інструментарій, що поєднує пошук релевантних джерел, аналіз вибраного контенту, прогноз його розвитку, і складається з математичних методик та технологічних компонувань даних у єдиний профіль для конкретної галузі за напрямом застосування. Також, досліджено комплексні методи розпізнавання фонетичного, синтаксичного і семантичного впливів інформаційних ресурсів на підсвідомість людини на основі технології аналізу фонетичної структури текстів, встановлення позитивного і негативного впливу окремих слів за їх звукокольоровими характеристиками [3], в інтересах розробки технологій інформаційної протидії. Інші напрацювання включають аналіз тексту шляхом визначення змістовних ознак його контенту і їх взаємозв'язків із завданнями ІАД [4].

Разом з цим, важливе місце у системі аналізу інформаційного потоку займає не лише розпізнавання змісту текстів, а й оцінювання їх емоційної забарвленості. Зокрема, інтерес представляє визначення методами машинного навчання емоційної забарвленості фрагменту тексту не лише за окремими словами, а за словосполученнями й цілими фразами. Щодо мови, значною поширеністю у світі відрізняється англійська мова, що є важливим чинником впливу на суспільну думку.

Тому **метою і основним змістом статті** є визначити найбільш оптимальний метод машинного навчання серед розглянутих для оцінювання емоційної забарвленості текстів англомовних ЗМІ.

Викладення основного матеріалу. Для вирішення цього завдання вивчались методи машинного навчання і їх властивості щодо опрацювання англомовних текстів. Набір для навчання класифікатора включає близько 1000 текстів з політично-спрямованих видань, розміщених в Інтернет, таких як The Guardian, The BBC, The Telegraph, Euronews, France24, Deutschewelle, TASS, Bloomberg, The CNN та ін.

В ході дослідження були розглянуті наступні методи машинного навчання для опрацювання текстів англомовних ЗМІ, з визначенням їх переваг та недоліків:

- наївний метод Байєса;
- метод логістичної регресії;
- метод штучної нейронної мережі.

Наївний метод Байєса. Першим методом машинного навчання, що розглядався, є наївний метод Байєса (наївний байєсівський класифікатор (НБК)).

Наївний байєсівський класифікатор (НБК) визначається двома способами. Перший – це поліноміальний метод Байеса, другий – це багатовимірна модель Бернуллі.

Поліноміальний метод Байеса або поліноміальна модель Naïve Bayes (NB), є імовірнісним методом навчання. Ця модель генерує один термін зі словника в кожній позиції документа.

Багатовимірна модель Бернуллі або модель Бернуллі, є альтернативою поліноміальній моделі. Вона еквівалентна незалежній бінарній моделі, яка генерує індикатор для кожного терміна словника: цифра 1 вказує на наявність терміна в документі, 0 вказує на його відсутність. Модель Бернуллі має ту ж часову складність, як і поліноміальна модель.

Різні породжувальні моделі передбачають різні стратегії оцінки та різні правила класифікації. Модель Бернуллі оцінює $P(t/d)$ як частку документів класу C, які містять терм t. На відміну від нього, поліноміальна модель оцінює $P(t/c)$ у вигляді частки лексем або частки позицій в документах класу C, які містять термін t.

На етапі класифікації тестового документу модель Бернуллі використовує двійкову інформацію про входження, ігноруючи кількість входжень, в той час як поліноміальна модель продовжує відслідковувати кількаразові входження. В результаті, модель Бернуллі, як правило, робить багато помилок при класифікації довгих документів.

Моделі також відрізняються у тому, як терми, котрі відсутні в документі, використовуються в класифікації. Вони не впливають на рішення класифікатора у поліноміальній моделі. Але у моделі Бернуллі імовірність невходження буде враховуватись при підрахунку $P(c|d)$, оскільки модель Бернуллі моделює відсутність термів явно.

Основними перевагами НБК є:

- простота реалізації.
- швидкий процес навчання. Обчислювальна складність навчання $O(|V|)$.
- не дивлячись на те, що припущення про незалежність класифікаційних ознак не є вірним в природній мові (значення слова залежать від контексту) НБК часто показує хороші результати при класифікації текстів.

Разом з цим, розглянутий метод має недоліки, які не дозволяють його використовувати у виконанні завдань оцінювання емоційної забарвленості інформаційного потоку з англомовних текстів ЗМІ. Адже, оскільки слова у природній мові не є незалежними, незважаючи на швидке навчання, НБК не є оптимальним алгоритмом для оцінювання емоційного забарвлення англомовних текстів ЗМІ, і не може бути достатньо продуктивним. Серед його основних недоліків – значення, які повертаються при класифікації, не можна трактувати, як імовірності, і це не дає можливості відповісти на питання, з якою часткою впевненості визначений клас [5].

Метод логістичної регресії. Наступним методом оцінювання текстів було розглянуто *логістичну регресію*. Це метод побудови лінійного класифікатору, який дозволяє оцінювати апостеріорні вірогідності належності об’єктів до класів. Так, якщо об’єкти описуються

числовими ознаками $f_j : X \to R, j = 1,...,n$ – тоді простір ознакових описів об'єктів є $X = R^n$, а Y – кінцева множина номерів (імен, міток) класів.

Навчальна вибірка задається парами "об'єкт", "відповідь".

$$X^m = \{(x_1 y_1),...,(x_m, y_m)\}.$$

Припускаючи, що $Y = \{-1,+1\}$, в логістичній регресії будується лінійний алгоритм класифікації $\alpha : X \to Y$ виду

$$\alpha(x, w) = sign(\sum_{j=1}^{n} w_j f_j(x) - w_0) = sign\langle x, w\rangle,$$

де w_j – вага *j-ї* ознаки, w_0 – поріг прийняття рішення.

$w = (w_0 w_{1,...}, w_n)$ – вектор ваг, $\langle x,w\rangle$ – скалярний добуток ознакового опису об'єкту на вектор ваг. Задача навчання лінійного класифікатору полягає у тому, щоб за вибіркою X^m налаштувати вектор ваг W. В логістичній регресії для цього вирішується задача мінімізації емпіричного ризику з функцією втрат спеціального виду. Після того, як рішення знайдено, стає можливим не лише вираховувати класифікацію для довільного об'єкту, але й оцінювати апостеріорні вірогідності його належності до класів [5, 6].

Дослідження методу логістичної регресії для оцінювання емоційної забарвленості інформаційного потоку англомовних текстів ЗМІ показало неможливість його використання для даної задачі. Внаслідок того, що цей метод вивчає лінійні взаємозв'язки між даними та результатом, а для вказаної задачі оцінювання текстів необхідно вивчати квадратичні зв'язки – кількість вхідних параметрів потрібно було б збільшити в квадратичне число разів. Тобто для вивчення квадратичного взаємозв'язку необхідно 10000 вхідних параметрів (кількість лексичних одиниць у тренувальному корпусі текстів) перетворити у 100000000 вхідних параметрів, що значно затягне тренування алгоритму.

Метод штучної нейронної мережі. В ході дослідження було встановлено, що для виконання завдань оцінювання емоційної забарвленості англомовних текстів доцільно використовувати підхід, який базується на методі Штучної нейронної мережі (ШНМ). Це математична модель та її програмне і апаратне застосування, побудовані за принципом організації та функціонування біологічних нейронних мереж – мереж нервових клітин живого організму [7, 8].

Алгоритм машинного навчання, покладений в основу роботи нейронної мережі, реалізований шляхом комбінування штучних нейронів (перцептронів), завдяки чому він може вирішувати важкі та комплексні задачі навчання зі вчителем, без вчителя, та навчання з підтримкою. Структура нейронної мережі може бути найрізноманітнішою, але її мінімальною одиницею є штучний нейрон.

Штучний нейрон має дві функції: суматорну та активаційну [9].

Суматорна (z) – являє собою добуток вхідного вектору або вихідних даних попереднього шару нейронної мережі на вагу перцептрону. Також до результату векторно додається коефіцієнт зміщення. Вага та коефіцієнт зміщення – це параметри нейронної мережі, які змінюються від ітерації до ітерації, і таким чином алгоритм доходить до свого глобального мінімуму.

Активаційна (a) – функція, що застосовується до результату суматорної функції і використовується для того, щоб проектувати результат на нелінійний простір. Таким чином, алгоритм може навчати різноманітні взаємозв'язки даних, а не тільки лінійні. Зазвичай активаційна функція може набувати наступних значень.

1. ReLU (Rectified linear unit) = max(0, z) [10];

2. Sigmoid = $\frac{1}{\left(1+e^{-z}\right)}$ [11];

3. Tanh = $\frac{\left(e^{z}-e^{-z}\right)}{\left(e^{z}+e^{-z}\right)}$ [12];

4. Linear = z.

В ході вивчення особливостей методів машинного навчання установлено, що найбільш використовуваною та результативною для виконання задачі є активаційна функція ReLU, вона і була застосована в більшості шарів нейронної мережі у нашому дослідженні. Сигмоїдна функція часто використовується в якості останньої активаційної функції, оскільки вона проектує результат на відрізок (0,1), та може інтерпретуватись, як імовірність виникнення події [9].

Комбінація таких штучних нейронів організує нейронну мережу, що здатна вивчати найбільш комплексні задачі. В результаті роботи алгоритму, ми отримуємо вихідний вектор, який і є індикатором точності роботи алгоритму, адже саме його зрівнюють з вектором істинних значень.

За різницею вихідного вектору та вектору істинних значень обчислюється функція втрат, яка є індикатором того, на скільки необхідно змінити параметри алгоритму, тобто ваги та коефіцієнти зміщення, що алгоритм працював краще. Зазвичай використовуються наступні функції втрат.

1. Середньоквадратична помилка:

$$\frac{1}{n} * \sqrt{(y-z)^2} \ ,$$

де y – вектор істинних значень,

z – вихідний вектор алгоритму,

n – потужність вектору, або кількість елементів в векторі [13].

2. Логістична помилка, або ентропія:

$$-\sum y * \log(z) \ \text{[14]}.$$

Середньоквадратична помилка використовується для задач регресії, тобто задачі прогнозування числа на нескінченному проміжку. Логістична помилка використовується для задач класифікації, тобто прогнозування на обмеженій кількості відповідей. Оскільки задача прогнозування емоції тексту, з обмеженою кількістю емоцій, – це задача класифікації, у задачі використовувалась логістична функція помилки.

Для того щоб оптимізувати роботу алгоритму, необхідно правильно змінювати параметри. Для цього слід використовувати алгоритм оптимізації: найбільш результативним та популярним з таких алгоритмів є градієнтний спуск [9]:

$$x^{(1)} = x^{(0)} - \gamma * \nabla F\left(x^{(0)}\right),$$

де γ – швидкість навчання,

∇F – градієнт функціоналу помилки.

Таким чином, параметри алгоритму змінюються відносно попередніх значень з урахуванням похідної функції втрат. Але дані значення можливо враховувати тільки для параметрів останнього шару нейронної мережі. Оскільки алгоритм є композицією функцій, ми можемо використати ланцюгове правило [15]:

$$\left(f \cdot g\right)' = \left(f' \cdot g\right) * g' .$$

Отже, за допомогою композиції елементарних функцій, з урахуванням функції помилки та поширенням цієї помилки через всю нейронну мережу, ми знаходимо оптимальні значення параметрів і алгоритм набуває здатності прогнозувати правильні значення для заданих вхідних векторів.

Окрім зазначеного вище, для оцінювання емоційної забарвленості англомовних текстів новин можна використовувати архітектуру рекурентних нейронних мереж (РНМ). Для цього пропонується використання такого методу ШНМ як Довга короткочасна пам'ять – ДКЧП. Це архітектура рекурентних нейронних мереж – РНМ, яка завдяки відносній нечутливості до довжини прогалин дає ДКЧП перевагу в численних застосуваннях над альтернативними РНМ, прихованими марковськими моделями та іншими методами навчання послідовностей [7, 9, 16, 17].

Отже, за допомогою композиції елементарних функцій, з урахуванням функції помилки та поширенням цієї функції через всю нейронну мережу, знаходяться оптимальні значення параметрів і алгоритм набуває здатності прогнозувати правильні значення для заданих вхідних векторів.

Слід зазначити, що аналіз роботи алгоритмів методів машинного навчання, досліджених у статті, проводився за допомогою методів метрики точності, дерева рішень, AUC-ROC та F-міри [18-20].

Висновки. Результати даного дослідження дають можливість вибрати оптимальний метод машинного навчання для аналізу новин англомовних ЗМІ, та виходячи з їх властивостей, визначати доцільність їх використання у низці завдань.

Результати, отримані в ході дослідження, свідчать про обмежену придатність більшості з розглянутих методів машинного навчання для визначення емоційної забарвленості текстів англомовних новин за напрямом даного дослідження. Так, аналіз методу НБК та перевірка його роботи за методами метрики точності тощо показав, що він не дає можливості встановити імовірність присутності визначеної емоції у кожному оцінюваному фрагменті. Аналіз методу логістичної регресії за вказаними вище методами в ході перевірки можливості використання за даним напрямом також показав низьку результативність, оскільки через необхідність вивчати квадратичні зв’язки в аналізі даного виду кількість вхідних параметрів потрібно було б збільшити в квадратичне число разів, що значно затягне навчання алгоритму.

Найбільш відповідним методом для оцінювання емоційного компоненту текстів англомовних новин, єдиним з методів машинного навчання, вивчених у даному дослідженні, є метод ШНМ. Зокрема, дієвими показали себе його складові: РНМ і ДКЧП. Результативною показала себе активаційна функція ReLU, що була використана у даному дослідженні в більшості шарів нейронної мережі. У якості останньої активаційної функції часто використовувалась сигмоїдна функція, яка при класифікації даних по відповідності певним емоціям у тексті дає можливість встановлювати імовірність приналежності до певного класу, проектуючи результат на відрізок (0,1), та інтерпретуючись як імовірність виникнення події. Переваги використання нейронних мереж, методу LSTM та функції ReLU дають можливість вивчати найрізноманітніші взаємозв’язки між даними, без необхідності значних витрат часу на роботу алгоритму.

Отримані висновки свідчать про важливість подальших досліджень за цим напрямом та пошуку нових можливостей, які дозволять поєднати обрані методи з іншими параметрами аналізу інформації. Результати наступних досліджень повинні бути враховані для інтеграції в систему діяльності інформаційно-аналітичних підрозділів.

ЛІТЕРАТУРА:

1. Писарчук О.О. Технологія автоматизованої багатовимірної оперативної та поглибленої аналітичної обробки актуальних інформаційних масивів / О.О. Писарчук, В.С. Косіков // Проблеми створення, випробування, застосування та експлуатації складних інформаційних систем. – 2015. – Вип. 10. – С. 183-195. // [Електр. Ресурс]. – Режим доступу: http://nbuv.gov.ua/UJRN/Psvz_2015_10_22.

2. Писарчук О.О. Методика прогнозування розвитку інформаційного контенту в мережі Інтернет / О. О. Писарчук, Д. В. Порада // Проблеми створення, випробування, застосування та експлуатації складних інформаційних систем. – 2015. – Вип. 10. – С. 170-182. // [Електр. Ресурс]. – Режим доступу: http://nbuv.gov.ua/UJRN/Psvz_2015_10_21.

3. Алімпієв А.М. Теоретичні основи створення технологій протидії прихованим інформаційним атакам в сучасній гібридній війні / А.М.Алімпієв, В.В. Бараннік, Т.В. Белікова, С.О. Сідченко//[Електр. Ресурс] – Режим доступу:
www.hups.mil.gov.ua/periodic-app/article/17669/soi_2017_4_26.pdf
4. Молодецька-Гринчук К.В. Метод виявлення ознак інформаційних впливів у соціальних інтернет-сервісах за змістовними ознаками. Радіоелектроніка, інформатика, управління. – 2017. – № 2. – С. 117-126. // [Електр. Ресурс]. – Режим доступу: http://www.irbis-nbuv.gov.ua/cgi-bin/irbis_nbuv/cgiirbis_64.exe?I21DBN=LINK&P21DBN=UJRN&Z21ID=&S21REF=10&S21CNR=20&S21STN=1&S21FMT=ASP_meta&C21COM=S&2_S21P03=FILA=&2_S21STR=riu_2017_2_15
5. Tom M. Mitchell. Generative and discriminative classifiers: Naive Bayes and Logistic regression. // [Електр. Ресурс] – Режим доступу: https://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf
6. Логистическая регрессия. Режим доступу:
http://www.machinelearning.ru/wiki/index.php?title=%D0%9B%D0%BE%D0%] // [Електр. Ресурс]
7. Haşim Sak, Andrew Senior, Francoise Beaufays. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. // [Електр. Ресурс] – Режим доступу: https://arxiv.org/abs/1402.1128
8. Искусственная_нейронная_сеть. // [Електр. Ресурс] – Режим доступу: https://ru.wikipedia.org/wiki/Искусственная_нейронная_сеть
9. Sebastian Ruder. An overview of gradient descent optimization algorithms. // [Електр. Ресурс] – Режим доступу https://arxiv.org/pdf/1609.04747.pdf
10. Xiaojie Jin, Chunyan Xu, Jiashi Feng, Yunchao Wei Junjun, Xiong Shuicheng Yan. Deep Learning with S-shaped rectified linear activation units. *12.2015.* // [Електр. Ресурс] – Режим доступу: https://arxiv.org/pdf/1512.07030.pdf
11. David Kriesel. A brief introduction to neural networks // [Електр. Ресурс] – Режим доступу: http://www.dkriesel.com/_media/science/neuronalenetze-en-zeta2-1col-dkrieselcom.pdf
12. Hyperbolic functions // [Електр. Ресурс] – Режим доступу: http://www.mathcentre.ac.uk/resources/workbooks/mathcentre/hyperbolicfunctions.pdf
13. Tianfeng Chai1 and Roland R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. // [Електр. Ресурс] – Режим доступу: https://www.geosci-model-dev-discuss.net/7/C473/2014/gmdd-7-C473-2014-supplement.pdf
14. Peter Sadowski Notes on Backpropagation // [Електр. Ресурс] – Режим доступу: https://www.ics.uci.edu/~pjsadows/notes.pdf
15. Differentiation: Chain Rule // [Електр. Ресурс] – Режим доступу: https://www.ucd.ie/t4cms/Differentiation - Chain Rule.pdf
16. Рекурентна нейронна мережа. // [Електр. Ресурс] – Режим доступу: https://uk.wikipedia.org/wiki/Рекурентна_нейронна_мережа
17. Прихована марковська модель // [Електр. Ресурс] – Режим доступу: https://uk.wikipedia.org/wiki/Прихована_марковська_модель
18. Václav Hlaváč. Classifier performance evaluation // [Електр. Ресурс] – Режим доступу: http://cmp.felk.cvut.cz/~hlavac/TeachPresEn/31PattRecog/13ClassifierPerformance.pdf
19. Analyzing machine-learning model performance // [Електр. Ресурс] – Режим доступу: https://github.com/IBM-Bluemix-Docs/knowledge-studio/blob/master/evaluate-ml.md
20. Classification: ROC and AUC // [Електр. Ресурс] – Режим доступу: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

REFERENCES:

1. Pisarchuk, O.O., Kosikov, V.S. (2015), Tehnologija avtomatizovanoї bagatovymirnoyi operatyvnoyi ta poglyblenoyi analitychnoyi obrobki aktual'nyh informaciynyh masyviv [the technique for automated multidimensional operational and profound analytical processing of actual information massifs], Problems of creation, testing, employment and maintenance of complex information systems, No.10, pp.183-195.
2. Pisarchuk, O. O., Porada D. V. (2015), Metodika prognozuvannja rozvytku informacijnogo kontentu v merezhi Internet [Methodology for prognosis of the information content development in the Internet], Problems of creation, testing, employment and maintenance of complex information systems, No.10, pp.170-182.
3. Alimpiєv, A.M., Barannik, V.V., Belikova, T.V., Sidchenko, S.O., (2017), “Teoretychni osnovy stvorennja tehnologij protydiї prihovanym informacijnym atakam v suchasnij gibrydnij vijni” [Theoretical

basis for development of technology of counteraction to implicit information attacks in the modern hybrid warfare]. [Online]. Available: www.hups.mil.gov.ua/periodic-app/article/17669/soi_2017_4_26.pdf

4. Molodets'ka-Grinchuk, K.V. (2017), "Metod vyjavlennja oznak informacijnyh vplyviv u social'nyh internet-servisah za zmistovnymy oznakami [The method for identification of information influence features in the social internet-services by content characteristics]. "Radioelectronics, Informatics, Control", No.2, pp.117-126.

5. Tom M. Mitchell, Hill McGraw. Generative and discriminative classifiers: Naive Bayes and Logistic regression. (2015), [Online]. Available: https://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

6. "Logisticheskaja regressija" [Logistic regression]. [Online]. Available: http://www.machinelearning.ru/wiki/index.php?title=%D0%9B%D0%BE%D0%]

7. Haşim Sak, Andrew Senior, Francoise Beaufays. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling, (2014). // [Online]. Available:
https://arxiv.org/abs/1402.1128

8. Iskusstvennaja_nejronnaja_set' [Artificial neural network]. [Online]. Available: https://ru.wikipedia.org/wiki/Iskusstvennaja_nejronnaja_set'

9. Sebastian Ruder. An overview of gradient descent optimization algorithms, (2017). [Online]. Available: https://arxiv.org/pdf/1609.04747.pdf

10. Xiaojie Jin, Chunyan Xu, Jiashi Feng, Yunchao Wei Junjun, Xiong Shuicheng Yan. Deep Learning with S-shaped rectified linear activation units. (2015). [Online]. Available: https://arxiv.org/pdf/1512.07030.pdf

11. David Kriesel. A brief introduction to neural networks, (2005). [Online]. Available: http://www.dkriesel.com/_media/science/neuronalenetze-en-zeta2-1col-dkrieselcom.pdf

12. Hyperbolic functions (2006). [Online]. Available: http://www.mathcentre.ac.uk/resources/workbooks/mathcentre/hyperbolicfunctions.pdf

13. Tianfeng Chai1, Roland R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, (2014). [Online]. Available: https://www.geosci-model-dev-discuss.net/7/C473/2014/gmdd-7-C473-2014-supplement.pdf

14. Peter Sadowski. Notes on Backpropagation. [Online]. Available: https://www.ics.uci.edu/~pjsadows/notes.pdf

15. Differentiation: Chain Rule. [Online]. Available: https://www.ucd.ie/t4cms/Differentiation - Chain Rule.pdf

16. Рекурентна нейронна мережа [Recurrent neural network]. [Online]. Available: https://uk.wikipedia.org/wiki/Рекурентна_нейронна_мережа

17. Прихована марковська модель [Hidden Markov model]. [Online]. Available: https://uk.wikipedia.org/wiki/Прихована_марковська_модель

18. Václav Hlaváč. Classifier performance evaluation. [Online]. Available:
http://cmp.felk.cvut.cz/~hlavac/TeachPresEn/31PattRecog/13ClassifierPerformance.pdf

19. Analyzing machine-learning model performance, (2017). [Online]. Available:
https://github.com/IBM-Bluemix-Docs/knowledge-studio/blob/master/evaluate-ml.md

20. Classification: ROC and AUC, (2018). [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

к.т.н. Бабіч О.М., Дух Д.І., Глухова А.С.

ОСОБЕННОСТИ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ОТНОСИТЕЛЬНО ИХ ИСПОЛЬЗОВАНИЯ В ПРОЦЕССЕ АНАЛИЗА АНГЛОЯЗЫЧНЫХ ИСТОЧНИКОВ

Исследованы методы машинного обучения для оценивания эмоциональной окраски текстов англоязычных средств массовой информации. Рассмотренны такие методы как наивный метод Байеса, логистическая регрессия, исскусственные нейронные сети. Определены их преимущества и недостатки относительно использования в указанной задаче. Так, несмотря на быстрое обучение, Наивный Байесовский классификатор в силу невозможности нахождения вероятностей значений при классификации, не является оптимальным алгоритмом для оценивания эмоциональной окраски англоязычных текстов СМИ и не может быть достаточно продуктивным. Метод логистической регрессии, в свою очередь, предусматривает изучение квадратических связей для выполнения задачи анализа эмоциональной окраски англоязычных текстов СМИ, что увеличит как количество параметров на входе, так и время обучения алгоритма. В ходе исследования преимущество отдано методу Искусственной нейронной сети,

алгоритмы которого дают возможность изучать разнообразные взаимосвязи между данными, при этом без необходимости в значительных временных затратах на работу алгоритма. Анализ работы алгоритмов рассмотренных в статье методов машинного обучения проводился с помощью методов метрики точності, дерева решений, AUC-ROC и F-мера.

***Ключевые слова:** информационно-аналитическая деятельность, машинное обучение, средства массовой информации, классификация, функция, алгоритм, нейронная сеть.*

Ph.D. Babich O.M., Dukh D.I., Gluhova A.S.

THE FEATURES OF MACHINE LEARNING METHODS IN REGARD TO THEIR EXERCISE IN THE PROCEDURE OF ENGLISH LANGUAGE SOURCES ANALYSIS

The modern information and analytical activity needs enhancement of data processing technology and impartial outputs. Dynamic and in-depth analysis of foreign language information array, in particular texts of English-language mass-media, is important for this activity performing. To analyze the informational stream, it is essential to evaluate the emotional coloring of text, not only to discern its content. Simultaneously, the text fragment should be evaluated not only by the separate words but also by the word combinations and entire phrases. Respectively, the research mission is to find the most optimal method for the emotional colouring evaluation of English mass media texts, among examined ones in this paper.

For this purpose, several Methods of machine learning are examined to evaluate emotional colouring of texts of English-written mass-media. Such methods as Naïve Bayes, Linear regression and Artificial neural network are considered. Their advantages and disadvantages are viewed to be exploited in the mentioned task. At this point Naïve Bayes classifier is not the proper algorithm for emotional colouring evaluation due to impossibility of finding probability of value. So it can`t be productive enough despite a rapid learning process. The Logistic regression method, in its turn, provides quadratic relations study to implement the task of emotional colouring analysis of English-written mass-media texts. This entails both entrance characteristics quantity and an algorithm learning time. The research ascertained that the technique of Artificial neural network is the most preferable out of viewed ones. Among them the algorithms of Gradient descent, the Recurrent neural network and the technique of Long short term memory are effective. The ANN algorithms make possible to study the vast scope of links between data, simultaneously without wasting much time for algorithm work. The defined techniques employment is able to enhance the outcome of the English language mass media texts analysis. Performance of machine-learning methods was analysed with the metrics of precision, AUC-ROC, F1 score and decision tree.

***Keywords:** information and analytical activity, mass media, machine learning, classification, function, algorithm, neural network.*